Express Mail Label No. EV402807440US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANT:	TAE-JIN AHN	)
		)
FOR:	APPARATUS AND METHOD FOR ENCODING DNA SEQUENCE, AND COMPUTER READABLE MEDIUM	)))

CLAIM FOR PRIORITY

Mail Stop Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

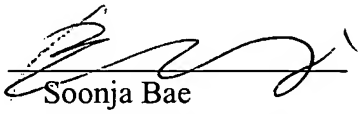
Dear Commissioner:

Enclosed herewith is a certified copy of Korean Patent Application No. 2003-0006543 filed on February 3, 2003. The enclosed Application is directed to the invention disclosed and claimed in the above-identified application.

Applicant hereby claims the benefit of the filing date of February 3, 2003, of the Korean Patent Application No. 2003-0006543, under provisions of 35 U.S.C. 119 and the International Convention for the protection of Industrial Property.

Respectfully submitted,

CANTOR COLBURN LLP

By: 

Soonja Bae
Reg. No. (See Attached)
Cantor Colburn LLP
55 Griffin Road South
Bloomfield, CT 06002
PTO Customer No. 23413
Telephone: (860) 286-2929
Fax: (860) 286-0115

Date: February 2, 2004

Express Mail Label No. EV402807440US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	TAE-JIN AHN	)
		)
For:	APPARATUS AND METHOD FOR ENCODING DNA SEQUENCE, AND COMPUTER READABLE MEDIUM	)

CLAIM FOR PRIORITY

Mail Stop Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

Applicant hereby claims the benefits of the filing date of January 30, 2004 to Korean Patent Application No. 2004-0005945 under provisions of 35 U.S.C. 119 and the International Convention for the protection of Industrial Property.

If any fees are due with regard to this claim for priority, please charge them to Deposit Account No. 06-1130 maintained by Applicant's attorneys.

Respectfully submitted,

CANTOR COLBURN LLP

By: _____
Soonja Bae
Reg. No. (SEE ATTACHED)
Cantor Colburn LLP
55 Griffin Road South
Bloomfield, CT 06002
PTO Customer No. 23413
Telephone: (860) 286-2929
Fax: (860) 286-0115

Date: February 2, 2004

대 한 민 국 특 허 청
KOREAN INTELLECTUAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Intellectual Property Office.

출 원 번 호 : 10-2003-0006543
Application Number

출 원 년 월 일 : 2003년 02월 03일
Date of Application FEB 03, 2003

출 원 인 : 삼성전자주식회사
Applicant(s) SAMSUNG ELECTRONICS CO., LTD.

2003 년 03 월 07 일

특 허 청
COMMISSIONER

【서지사항】

【서류명】	특허출원서
【권리구분】	특허
【수신처】	특허청장
【참조번호】	0004
【제출일자】	2003.02.03
【국제특허분류】	G03M
【발명의 명칭】	ＤＮＡ 서열 부호화 장치 및 방법
【발명의 영문명칭】	Apparatus for encording DNA sequence and method of the same
【출원인】	
【명칭】	삼성전자 주식회사
【출원인코드】	1-1998-104271-3
【대리인】	
【성명】	이영필
【대리인코드】	9-1998-000334-6
【포괄위임등록번호】	2003-003435-0
【대리인】	
【성명】	이해영
【대리인코드】	9-1999-000227-4
【포괄위임등록번호】	2003-003436-7
【발명자】	
【성명의 국문표기】	안태진
【성명의 영문표기】	AHN,Tae Jin
【주민등록번호】	760406-1018324
【우편번호】	135-270
【주소】	서울특별시 강남구 도곡동 951-7 낙연주택 201호
【국적】	KR
【심사청구】	청구
【취지】	특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 출원심사 를 청구합니다. 대리인 이영필 (인) 대리인 이해영 (인)

【수수료】

【기본출원료】	20	면	29,000 원
【가산출원료】	5	면	5,000 원
【우선권주장료】	0	건	0 원
【심사청구료】	12	항	493,000 원
【합계】	527,000 원		

【첨부서류】 1. 요약서·명세서(도면)_1통

【요약서】

【요약】

DNA 서열 부호화 장치 및 방법이 개시된다. 비교부는 DNA 정보가 알려진 원본서열과 코딩될 대상서열이 최대한 일치하도록 정렬한 후 차이점을 비교한다. 분할부는 정렬된 원본서열과 대상서열을 일정한 크기로 분할한다. 변환부는 원본서열과 대상서열의 차이점을 소정 개수의 문자를 이용하여 문자열로 변환한다. 코딩부는 코드저장부에 저장되어 있는 일정 크기의 코드를 이용하여 문자열을 구성하는 각각의 문자를 코딩한다. 압축부는 코딩결과를 통상적인 압축방법을 이용하여 압축하며, 압축된 결과는 서열저장부에 저장된다. 본 발명에 따르면, 정보의 손실없이 높은 압축율로 DNA 서열을 압축하여 저장할 수 있으며, 데이터의 전송속도 및 검색효율을 높일 수 있다.

【대표도】

도 1

【색인어】

게놈, DNA 서열, 압축, 부호화

【명세서】

【발명의 명칭】

DNA 서열 부호화 장치 및 방법{Apparatus for encording DNA sequence and method of the same}

【도면의 간단한 설명】

도 1은 본 발명에 따른 DNA 서열 부호화 장치에 대한 일실시예의 구성을 도시한 블록도,

도 2는 서열비교의 일 예로 NCBI에서 제공하는 기본 툴인 genomics blast를 사용하여 원본서열과 대상서열을 비교한 결과를 도시한 도면,

도 3은 비교부에서 정렬된 원본서열과 대상서열의 차이점을 문자열로 변환하는 원리를 도시한 도면,

도 4는 문자열을 코드화하기 위한 4비트 코드의 일예를 도시한 도면,

도 5는 mody3 유전자의 엡손영역을 문자열로 변환한 결과 및 문자열을 4비트의 코드로 코드화한 결과를 도시한 도면, 그리고,

도 6은 본 발명에 따른 DNA 서열 부호화 방법에 대한 일 실시예의 수행과정을 도시한 흐름도이다.

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<7> 본 발명은 DNA 서열 부호화 장치 및 방법에 관한 것으로, 보다 상세하게는, 보다 효율적인 압축을 통해 저장공간 및 전송 트래픽의 용량을 줄일 수 있도록 DNA 서열을 부호화하는 장치 및 방법에 관한 것이다.

<8> 생명공학이 발달함에 따라 개체의 유전정보를 구성하는 DNA 서열이 밝혀지고 있다. 이러한 DNA 서열에 대한 연구결과는 개체의 형질변환, 질병추적 등 다양한 분야에 적용되며, 이를 위해 DNA 서열을 컴퓨터에서 이용할 수 있는 형태로 변환하여 저장할 필요가 있다. 그러나 DNA 서열은 정보량이 상당히 큰 관계로 상당한 저장비용이 소요된다. 따라서, DNA 서열의 저장, 전송, 검색 등을 위해 DNA 서열을 압축하는 것이 요구된다.

<9> DNA 서열의 압축방법에는 크게 사전기반압축방법(dictionary based compression method)과 비사전기반압축방법(non-dictionary based compression method)이 있다. 이중에서 사전기반압축방법의 압축율이 높으며, 일반적인 환경하에서 70-80%의 압축율을 보인다. 그러나, 이러한 압축기술은 게놈 전체 영역에 이르는 크기의 DNA 서열에 적용하기 어려운 단점이 있다.

<10> 최근 발표된 DNA 서열의 압축기술 중 성능이 가장 양호한 기술은 전체 게놈을 압축하는 데에도 적용할 수 있다. 이에 의하면, 일반적인 환경하에서 70-80%의 압축율을 보장할 수 있고, e-coli 게놈의 경우 96.6%의 압축율을 보장하는 것으로 발표되었다. 그러

나, 이러한 압축율은 단순한 추정치일 뿐 이러한 압축율을 달성하기 위한 구체적인 구성이 제시되지는 않은 상태이다.

【발명이 이루고자 하는 기술적 과제】

<11> 본 발명이 이루고자 하는 기술적 과제는, 효율적인 압축을 통해 저장공간 및 전송 트래픽의 용량을 줄일 수 있도록 DNA 서열을 부호화하는 장치 및 방법을 제공하는 데 있다.

【발명의 구성 및 작용】

<12> 상기의 기술적 과제를 달성하기 위한, 본 발명에 따른 DNA 서열 부호화 장치는, DNA 정보가 알려진 원본서열을 기준으로 부호화할 대상서열을 정렬하고 상기 원본서열과 상기 대상서열의 차이점을 비교하는 비교부; 상기 정렬된 원본서열과 대상서열을 소정의 크기로 분할하는 분할부; 상기 비교부가 추출한 원본서열과 대상서열의 차이점을 소정 개수의 문자에 의해 문자열로 변환하는 변환부; 소정 크기의 변환코드가 저장되는 코드저장부; 및 상기 문자열을 구성하는 각각의 문자를 상기 변환코드에 의해 코딩하는 코딩부;를 구비한다.

<13> 상기 문자는 DNA를 구성하는 염기를 나타내는 4개의 제1문자, 0 에서 9까지의 9개의 제2문자, 상기 차이점의 시작 및 종료를 나타내는 1개의 제3문자, 및 상기 차이점의 연속여부를 나타내는 1개의 제4문자로 구성되는 것이 바람직하다.

<14> 바람직하게는, 상기 변환부는 상기 차이점 각각에 대해 상기 차이점의 시작, 상기 차이점의 시작위치, 상기 차이점의 연속여부, 상기 차이점을 구성하는 염기가 연속되는 베이스의 개수, 상기 차이점을 구성하는 염기, 상기 차이점의 종료, 및 상기 차이점의

시작위치로부터 상기 차이점의 종료위치까지의 거리를 각각 상기 제3문자, 상기 제2문자, 상기 제4문자, 상기 제2문자, 상기 제1문자, 상기 제3문자, 및 상기 제2문자로 변환하고 변환된 문자가 연속적으로 배열된 상기 문자열을 출력한다.

<15> 바람직하게는, 상기 차이점의 형태는 상기 원본서열과 상기 대상서열의 시작영역이 불일치하는 시작영역불일치, 상기 원본서열에는 존재하는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하지 않음을 나타내는 공백, 상기 원본서열과 상기 대상서열의 대응되는 하나의 베이스 위치에 상이한 염기가 존재하는 단일베이스쌍불일치, 상기 원본서열에는 존재하지 않는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하는 삽입, 상기 원본서열과 상기 대상서열의 대응되는 복수의 베이스 위치에 상이한 염기가 존재하는 다중베이스쌍불일치, 및 상기 원본서열과 상기 대상서열의 종료영역이 불일치하는 종료영역불일치를 포함한다.

<16> 상기 변환코드는 상기 소정 개수의 문자 각각에 대응되는 4비트의 코드인 것이 바람직하다.

<17> 바람직하게는, 상기 변환코드에 의해 코딩된 상기 대상서열을 압축하는 압축부; 및 상기 압축된 대상서열이 저장되는 서열저장부;를 더 구비한다.

<18> 상기의 다른 기술적 과제를 달성하기 위한, 본 발명에 따른 DNA 서열 부호화 방법은, DNA 정보가 알려진 원본서열을 기준으로 부호화할 대상서열을 정렬하는 단계; 상기 원본서열과 상기 대상서열의 차이점을 비교하는 단계; 상기 정렬된 원본서열과 대상서열을 소정의 크기로 분할하는 단계; 상기 원본서열과 대상서열의 차이점을 소정 개수의 문자에 의해 문자열로 변환하는 단계; 및 상기 문자열을 구성하는 각각의 문자를 소정 개수의 변환코드에 의해 코딩하는 단계;를 포함한다.

<19> 상기 문자는 DNA를 구성하는 염기를 나타내는 4개의 제1문자, 0 에서 9까지의 9개의 제2문자, 상기 차이점의 시작 및 종료를 나타내는 1개의 제3문자, 및 상기 차이점의 연속여부를 나타내는 1개의 제4문자로 구성되는 것이 바람직하다.

<20> 바람직하게는, 상기 변환단계는, 상기 차이점 각각에 대해 상기 차이점의 시작을 나타내는 상기 제3문자를 부여하는 단계; 상기 차이점의 시작위치를 나타내는 상기 제2문자를 부여하는 단계; 상기 차이점의 연속여부를 나타내는 상기 제4문자를 부여하는 단계; 상기 차이점을 구성하는 염기가 연속되는 베이스의 개수를 나타내는 상기 제2문자를 부여하는 단계; 상기 차이점을 구성하는 염기를 나타내는 상기 제1문자를 부여하는 단계; 상기 차이점의 종료를 나타내는 상기 제3문자를 부여하는 단계; 상기 차이점의 시작위치로부터 상기 차이점의 종료위치까지의 거리를 나타내는 상기 제2문자를 부여하는 단계; 및 상기 부여된 문자가 연속적으로 배열된 상기 문자열을 출력하는 단계;를 포함한다.

<21> 바람직하게는, 상기 차이점의 형태는 상기 원본서열과 상기 대상서열의 시작영역이 불일치하는 시작영역불일치, 상기 원본서열에는 존재하는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하지 않음을 나타내는 공백, 상기 원본서열과 상기 대상서열의 대응되는 하나의 베이스 위치에 상이한 염기가 존재하는 단일베이스쌍불일치, 상기 원본서열에는 존재하지 않는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하는 삽입, 상기 원본서열과 상기 대상서열의 대응되는 복수의 베이스 위치에 상이한 염기가 존재하는 다중베이스쌍불일치, 및 상기 원본서열과 상기 대상서열의 종료영역이 불일치하는 종료영역불일치를 포함한다.

<22> 상기 변환코드는 상기 소정 개수의 문자 각각에 대응되는 4비트의 코드인 것이 바람직하다.

<23> 바람직하게는, 상기 변환코드에 의해 코딩된 상기 대상서열을 압축하는 단계; 및 상기 압축된 대상서열을 저장하는 단계;를 더 포함한다.

<24> 이에 의해, DNA 서열을 90%이상의 압축효율로 정보의 손실없이 압축하여 저장할 수 있다. 또한, 높은 효율로 DNA 서열을 압축할 수 있으므로, 게놈서열이나 게놈의 특정영역에 대한 다수의 DNA 서열을 저장하는 데 이용될 수 있다.

<25> 이하에서, 첨부된 도면들을 참조하여 본 발명에 따른 DNA 서열 부호화 장치 및 방법의 바람직한 실시예에 대해 상세하게 설명한다.

<26> 도 1은 본 발명에 따른 DNA 서열 부호화 장치에 대한 일실시예의 구성을 도시한 블록도이다.

<27> 도 1을 참조하면, 본 발명에 따른 DNA 서열 부호화 장치(100)는, 비교부(110), 분할부(120), 변환부(130), 코딩부(140), 압축부(150), 코드저장부(160), 및 서열저장부(170)를 갖는다.

<28> 비교부(110)는 DNA 정보가 알려진 원본서열을 기준으로 코딩될 대상서열을 정렬한 후 차이점을 비교한다. 이 때, 비교부(110)는 원본서열과 대상서열이 최대한 일치하도록 정렬한다. 분할부(120)는 정렬된 원본서열과 대상서열을 일정한 크기로 분할한다. 이러한 분할은 서열저장부(170)의 전체 용량의 15%의 크기로 수행되는 것이 바람직하다. 도 2에는 NCBI에서 제공하는 기본 툴인 genomics blast를 사용하여 원본서열과 대상서열을 비교한 결과가 도시되어 있다. 비교결과는 text, html, xml 등과 같은 형식의 문서로 출

력될 수 있다. 또한, 공지의 파싱기법을 이용하면 비교결과로부터 원본서열과 대상서열의 차이점만을 추출할 수 있다.

<29> 변환부(130)는 비교부(110)에서 추출된 원본서열과 대상서열의 차이점을 16개의 문자를 이용하여 문자열로 변환한다. 원본서열과 대상서열을 정렬할 때 나타날 수 있는 서열의 차이점은 여섯가지 패턴으로 구분될 수 있다. 변환부(130)는 이러한 여섯가지 패턴을 16개의 문자를 사용하여 원본서열과 대성서열의 차이점을 문자열로 표현한다. 16개의 문자는 숫자 10가지, DNA 서열의 종류를 표시하는 형태식별자 4가지, 정보간의 구분을 위한 문자 2가지로 구성된다. 표 1에는 원본서열과 대상서열의 차이점을 표현하는 16개의 문자 및 설명이 기재되어 있다.

<30> 【표 1】

문자	설명	
A	adenine	차이가 생기는 부분의 다른 DNA 염기 코드
T	thymine	
G	guanine	
C	cytocine	
0 ~ 9	차이점의 위치, 차이점의 연속된 길이, 차이점의 마지막 위치까지의 거리	
/	차이점의 기록 시작 및 종료	
~	차이점의 연속을 나타내는 구분자	

<31> 이하에서, 도 3에 도시된 원본서열과 대상서열을 예로들어 차이점을 문자열로 변환하는 원리에 대해 설명한다. 아래에 제시된 변환원리는 하나의 예이며 본 발명의 사상을 해하지 않는 범위에서 다양한 방법이 채택될 수 있음은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 자명한 사실이다.

<32> 먼저 차이점에 대한 패턴을 분석한다.

<33> A. 시작영역불일치: X_{-3}에서 X_{-1}까지의 영역은 시작영역이 불일치하는 영역으로 원본서열에는 존재하지 않는 gac가 대상서열에 존재한다.

<34> B. 공백: X_6에서 X_7까지의 영역은 대상서열에 염기가 존재하지 않는 영역으로 원본서열에 존재하는 ta가 대상서열에는 존재하지 않는다.

<35> C. 단일베이스쌍불일치: X_{11}은 원본서열과 대상서열의 염기가 일치하지 않는 지점이다.

<36> D. 삽입: X13과 X14 사이에 원본서열에는 존재하지 않는 atgcat가 대상서열에 존재한다.

<37> E. 다중베이스쌍불일치: X16에서 X18까지의 영역은 복수개의 베이스에 걸쳐 원본서열과 대상서열의 염기가 일치하지 않는 영역이다.

<38> F. 종료영역불일치: X23에서 X24까지의 영역은 시작영역이 불일치하는 영역으로 원본서열에는 존재하지 않는 ag가 대상서열에 존재한다.

<39> 다음으로, 차이점에 대한 패턴을 순차적으로 문자로 변환한다.

<40> 패턴 A를 문자열로 변환하면 "/-3~3gac/3"이다. 여기에서, "/"은 차이점 기록의 시작을 나타내는 문자이고, "-3"은 차이점이 시작되는 위치를 나타내는 문자로 X_0를 기준으로 음의 방향으로 3만큼 이동한 위치로부터 대상서열의 염기가 존재함을 나타내고, "~"은 차이점이 연속됨을 나타내는 문자이고, "3"은 차이점의 연속된 길이를 나타내는 문자이고, "gac"는 차이가 생기는 부분의 DNA 서열을 나타내는 문자이고, "/"은 차이점 기록의 종료를 나타내는 문자이며, "3"은 차이점의 시작위치로부터 차이점의 마지막위치까지의 거리를 나타내는 문자이다.

<41> 패턴 B를 문자열로 변환하면 "/6/2"이다. 여기에서, "6"은 패턴 A의 차이점의 시작위치로부터 차이점의 마지막위치까지의 거리를 나타내는 "3"에 의해 결정된 위치인 X_0로부터 양의 방향으로 6개 이동한 위치인 X_6으로부터 차이점이 시작됨을 의미한다. 또한, "2"는 차이점의 시작위치인 X_6으로부터 차이점의 마지막위치까지의 거리를 나타낸다.

<42> 패턴 C를 문자열로 변환하면 "/3~1c/1"이다. 여기에서, "3"은 패턴 B의 차이점의 시작위치로부터 차이점의 마지막위치까지의 거리를 나타내는 "2"에 의해 결정된 위치인 X_8로부터 양의 방향으로 3개 이동한 위치인 X_{11}로부터 차이점이 시작됨을 의미한다. 또한, "~1"은 연속되는 베이스의 개수가 1개임을 의미하며, "c"는 차이가 있는 DNA 서열을 나타내고, "1"은 차이점의 시작위치인 X_{11}로부터 차이점의 마지막위치까지의 거리를 나타낸다.

<43> 패턴 D를 문자열로 변환하면 "/1~6atgcat/1"이다. 여기에서, "1"은 패턴 C의 차이점의 시작위치로부터 차이점의 마지막위치까지의 거리를 나타내는 "1"에 의해 결정된 위치인 X_{12}로부터 양의 방향으로 1개 이동한 위치인 X_{13}으로부터 차이점이 시작됨을 의미한다. 또한, "~6"은 연속되는 베이스의 개수가 6개임을 의미하며, "atgcat"는 차이가 있는 DNA 서열을 나타낸다. 또한, "1"은 차이점의 시작위치인 X_{13}으로부터 차이점의 마지막위치까지의 거리를 나타내며, 거리가 "1"이므로 삽입임을 알 수 있다.

<44> 패턴 E를 문자열로 변환하면 "/2~3tcc/3"이다. 여기에서, "2"는 패턴 D의 차이점의 시작위치로부터 차이점의 마지막위치까지의 거리를 나타내는 "1"에 의해 결정된 위치인 X_{14}로부터 양의 방향으로 2개 이동한 위치인 X_{16}으로부터 차이점이 시작됨을 의미한다. 또한, "~3"은 연속되는 베이스의 개수가 3개임을 의미하며, "tcc"는 차이가 있는 DNA 서

열을 나타낸다. 또한, "3"은 차이점의 시작위치인 X_{16}으로부터 차이점의 마지막위치까지의 거리를 나타낸다.

<45> 패턴 F를 문자열로 변환하면 "/4~2ag/2"이다. 여기에서, "4"는 패턴 E의 차이점의 시작위치로부터 차이점의 마지막위치까지의 거리를 나타내는 "3"에 의해 결정된 위치인 X_{19}로부터 양의 방향으로 4개 이동한 위치인 X_{22}로부터 차이점이 시작됨을 의미한다. 또한, "~2"은 연속되는 베이스의 개수가 2개임을 의미하며, "ag"는 차이가 있는 DNA 서열을 나타낸다. 또한, "2"는 차이점의 시작위치인 X_{22}로부터 차이점의 마지막위치까지의 거리를 나타낸다.

<46> 이상의 결과를 하나의 문자열로 나타내면 다음과 같으며 문자 하나가 1바이트이므로 총 50바이트의 크기를 갖는다.

<47> "/-3~3gac/3/6/2/3~1c/1/1~6atgcat/1/2~3tcc/3/4~2ag/2"

<48> 코딩부(140)는 코드저장부(160)에 저장되어 있는 4비트 크기의 코드를 이용하여 문자열을 구성하는 각각의 문자를 코딩한다. 코드저장부(160)에 저장되어 있는 코드의 일예가 도 4에 도시되어 있다. 도 3에 도시된 각각의 패턴에 대한 문자열을 도 4에 도시된 코드를 이용하여 코딩된 결과는 다음과 같다.

<49> /-3~3gac/3: 11100000000000111111001111001010110111100011

<50> /6/2: 1110011011100010

<51> /3~1c/1: 11100011111100011101111000011

<52> /1~6atgcat/1: 11100110111110101011110011011010110111100001

<53> /2~3tcc/3: 11100010111100111011110111011110011

<54> /4~2ag/2: 11100100111100101010110011100010

<55> 따라서, 코딩부(140)에서 출력되는 최종적인 코딩결과는

11100000000000111111001111001010110111100011111001101110001011100011111100011101111000011110011011111010101111100110110101101111000011110001011110011101111101110111100011111001001111001010101100111000 10이며, 크기는 25바이트이다.

<56> 압축부(150)는 코딩결과를 통상적인 압축방법을 이용하여 압축한다. 압축된 결과는 서열저장부(170)에 저장된다.

<57> 원본서열과 대상서열의 차이점을 문자열로 변환한 후 4비트의 코드에 의해 코드화하는 과정을 mody3 유전자의 엡손(exon)영역에 적용하면 98.9%이상의 압축율을 얻을 수 있다. 또한, 코드화된 mody3 유전자의 엡손영역을 압축하면 보다 높은 압축율이 얻어진다. 도 5에는 mody3 유전자의 엡손영역을 문자열로 변환한 결과 및 문자열을 4비트의 코드로 코드화한 결과가 도시되어 있다. 도 5를 참조하면, 5552바이트의 크기를 갖는 유전자의 엡손영역이 122바이트의 문자열로 변환된 후 61바이트의 코드열로 코드화되며, 압축율은 98.9%임을 알 수 있다.

<58> 본 발명에 따른 DNA 서열 부호화 방법은 생물정보(bioinformatics)연구를 위한 통상적인 계산 장치인 PC, 워크스테이션, 슈퍼 컴퓨터 등에서 구현될 수 있다. 게놈 서열이 알려진 생물 개체에 대한 DNA 서열의 부호화 과정과 압축 과정은 여섯 단계로 구분할 수 있다.

<59> 도 6은 본 발명에 따른 DNA 서열 부호화 방법에 대한 일 실시예의 수행과정을 도시한 흐름도이다.

<60> 도 6을 참조하면, 밝혀진 게놈 서열과 저장할 생물 개체의 서열의 차이점을 추출한다(S600). S600단계에서 서열의 비교는 생물정보학분야에서 널리 알려진 통상적인 비교방법을 이용하여 수행될 수 있다. 본 발명에서 사용될 수 있는 서열비교방법에는 Blast, Blat, Fasta, Smith Wateman Algorithm 등이 있다. 이러한 방법을 이용하여 서열을 정렬·비교하고 결과파일을 공지의 파싱기술에 의해 파싱하여 차이점을 얻는다. 본 발명의 목표는 두 DNA 서열의 차이점만을 부호화하는 것이므로 DNA 서열의 정렬·비교의 목표는 두 DNA 서열이 최대한 일치하도록 하는 것이다.

<61> 다음으로, S600단계를 수행하여 얻은 결과를 메모리에서 처리하기에 적합한 크기로 분할한다(S610). 게놈 서열 전체는 수백 메가의 크기를 갖기 때문에 결과파일 전체에 대해 코딩하는 것은 바람직하지 않다. 따라서, 비교·정렬결과를 본 발명에 따른 DNA 서열 코딩장치에 구비된 전체 메모리의 15%에 해당하는 크기로 분할한다.

<62> 다음으로, 원본서열과 대상서열의 차이점을 문자열로 변환한다(S620). 원본서열과 대상서열을 정렬할 때 나타날 수 있는 서열의 차이점은 여섯가지 패턴으로 구분될 수 있다. S620단계에서는 이러한 여섯가지 패턴을 16개의 문자를 사용하여차이점을 문자열로 변환한다. 16개의 문자는 숫자 10가지, DNA 서열의 종류를 표시하는 형태식별자 4가지, 정보간의 구분을 위한 문자 2가지로 구성된다.

<63> 서열의 차이점의 패턴은 시작영역불일치(Start region mismatch), 공백영역(Blank region), 단일베이스쌍불일치(Single base pair mismatch), 다중베이스쌍불일치(Multiple base pair mismatch), 삽입영역(Inserted region), 및 종료영역불일치(End region mismatch)와 같이 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 용이하게 알 수 있는 용어들로 표현될 수 있다.

<64> 위에서 제시한 16개의 문자를 조합하면 서열의 차이점의 6가지 패턴에 대해 차이점이 존재하는 영역의 위치, 차이가 나는 DNA 서열, 차이의 길이정보 등을 문자열로 표현할 수 있다. 문자열로 표현된 서열은 비교의 대상이 되었던 서열과의 대조에 의해 원래의 서열로 정보의 손실없이 복원될 수 있다. 이러한 복원과정은 DNA 서열을 문자열로 표현하는 과정을 역으로 적용하여 수행된다.

<65> 다음으로, 문자열로 표현된 DNA 서열을 4비트의 코드에 의해 부호화한다(S630). 서열의 차이점을 16개의 문자에 의해 표현한 문자열을 구성하는 각각의 문자는 4비트의 코드로 나타낼 수 있다.

<66> 다음으로, 부호화된 결과를 통상의 압축 알고리즘을 사용하여 압축한다(S640). 본 발명에서 사용될 수 있는 압축 알고리즘은 LZ78, 호프만 코딩, 산술코딩 등과 같이 데이터 압축분야에서 널리 알려진 기술을 구현한 툴이 될 수 있다. 나아가, 유전정보의 압축과 관련된 다양한 공지의 압축기술이 사용될 수 있다. 압축된 DNA 서열은 하드디스크, CD, 등과 같은 다양한 저장수단에 저장된다(S650).

<67> 본 발명은 대상서열을 이미 알려진 원본서열과 비교하여 차이점만을 부호화하여 압축하므로 원본서열과의 상동성이 압축효율을 결정한다. 또한, 일반적인 생물학적 지식에 의하면 같은 종 내에서 DNA 서열의 동일성은 99% 이상이므로, 1% 이하의 차이점만이 기록의 대상이 된다고 할 수 있다. 따라서, 본 발명을 인간 게놈서열의 압축및 저장에 적용하면 98.65% 이상의 압축율을 기대할 수 있다.

<68> 이것은 다음의 조건에서 설명되며, 이러한 가정은 본 발명이 속한 기술 분야에 익숙한 사람이 충분히 받아들일 수 있는 정도의 것이다. 일반적으로 결핍이나 삽입에 의한 차이는 거의 일어나지 않으므로 모든 차이점이 단일베이스쌍불일치라 가정하고, 일반적

인 유전학의 가설에 따라 100개의 bp(base pair)마다 하나씩의 차이점이 생길 경우 기록할 양은 원래 정보의 양의 1%가 된다. 따라서, 전체의 1%가 부호화되어야 하며 문자열로 변환하는 과정에서 각각의 bp당 8글자(/100~1/1)가 더 기록되어야 하므로 8%의 기록양이 증가한다. 결과적으로 기록할 정보의 양은 원래 정보의 양의 9%가 된다. 그러나, 문자열을 4비트의 코드로 표현하면 기록할 정보의 양은 반으로 줄어들게 되며, 70%의 압축율을 가진 압축 알고리즘에 의해 압축하면 최종적으로 기록할 정보의 양은 원래 정보의 양의 1.35%가 된다. 따라서 인간 유전체 전체를 압축할 경우 이론상 보장할 수 있는 최저 압축율은 98.65% 이상이라고 할 수 있다.

<69> 이상에서 본 발명의 바람직한 실시예에 대해 도시하고 설명하였으나, 본 발명은 상술한 특정의 바람직한 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변형 실시가 가능한 것은 물론이고, 그와 같은 변경은 청구범위 기재의 범위 내에 있게 된다.

【발명의 효과】

<70> 본 발명에 따른 DNA 서열 부호화 장치 및 방법에 의하면 90%이상의 압축효율로 정보의 손실없이 DNA 서열을 압축하여 저장할 수 있다. 또한, 높은 효율로 DNA 서열을 압축할 수 있으므로, 게놈서열이나 게놈의 특정영역에 대한 다수의 DNA 서열을 저장하는데 이용될 수 있다. 일예로, 특정 질환을 일으키는 유전자가 발견되어 만명의 환자에 대하여 그 유전자의 서열을 밝혀내고 저장할 경우에 데이터를 압축하여 저장함으로써 저장공간을 줄일 수 있다. 나아가, 데이터의 전송속도 및 검색효율을 높일 수 있다. 또한, DNA 서열의 차이점만을 기록하므로 서로 다른 DNA 서열의 효율적인 비교·검색에 응용될

수 있다. 예를 들어, 특정질환은 일으키는 유전자에 대하여 만명의 환자와 정싱인의 DNA 서열이 존재할 때, 만명의 환자들과 정상인, 정상인과 정상인의 차이를 보이는 서열을 효율적으로 검색할 수 있다.

【특허청구범위】

【청구항 1】

DNA 정보가 알려진 원본서열을 기준으로 부호화할 대상서열을 정렬하고 상기 원본서열과 상기 대상서열의 차이점을 비교하는 비교부;

상기 정렬된 원본서열과 대상서열을 소정의 크기로 분할하는 분할부;

상기 비교부가 추출한 원본서열과 대상서열의 차이점을 소정 개수의 문자에 의해 문자열로 변환하는 변환부;

소정 크기의 변환코드가 저장되는 코드저장부; 및

상기 문자열을 구성하는 각각의 문자를 상기 변환코드에 의해 코딩하는 코딩부;를 포함하는 것을 특징으로 하는 DNA 서열 부호화 장치.

【청구항 2】

제 1항에 있어서,

상기 문자는 DNA를 구성하는 염기를 나타내는 4개의 제1문자, 0 에서 9까지의 9개의 제2문자, 상기 차이점의 시작 및 종료를 나타내는 1개의 제3문자, 및 상기 차이점의 연속여부를 나타내는 1개의 제4문자로 구성되는 것을 특징으로 하는 DNA 서열 부호화 장치.

【청구항 3】

제 2항에 있어서,

상기 변환부는 상기 차이점 각각에 대해 상기 차이점의 시작, 상기 차이점의 시작위치, 상기 차이점의 연속여부, 상기 차이점을 구성하는 염기가 연속되는 베이스의

개수, 상기 차이점을 구성하는 염기, 상기 차이점의 종료, 및 상기 차이점의 시작위치로부터 상기 차이점의 종료위치까지의 거리를 각각 상기 제3문자, 상기 제2문자, 상기 제4문자, 상기 제2문자, 상기 제1문자, 상기 제3문자, 및 상기 제2문자로 변환하고 변환된 문자가 연속적으로 배열된 상기 문자열을 출력하는 것을 특징으로 하는 DNA 서열 부호화 장치.

【청구항 4】

제 1항에 있어서,

상기 차이점의 형태는 상기 원본서열과 상기 대상서열의 시작영역이 불일치하는 시작영역불일치, 상기 원본서열에는 존재하는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하지 않음을 나타내는 공백, 상기 원본서열과 상기 대상서열의 대응되는 하나의 베이스 위치에 상이한 염기가 존재하는 단일베이스쌍불일치, 상기 원본서열에는 존재하지 않는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하는 삽입, 상기 원본서열과 상기 대상서열의 대응되는 복수의 베이스 위치에 상이한 염기가 존재하는 다중베이스쌍불일치, 및 상기 원본서열과 상기 대상서열의 종료영역이 불일치하는 종료영역불일치를 포함하는 것을 특징으로 하는 DNA 서열 부호화 장치.

【청구항 5】

제 1항에 있어서,

상기 변환코드는 상기 소정 개수의 문자 각각에 대응되는 4비트의 코드인 것을 특징으로 하는 DNA 서열 부호화 장치.

【청구항 6】

제 1항에 있어서,

상기 변환코드에 의해 코딩된 상기 대상서열을 압축하는 압축부; 및

상기 압축된 대상서열이 저장되는 서열저장부;를 더 포함하는 것을 특징으로 하는 DNA 서열 부호화 장치.

【청구항 7】

DNA 정보가 알려진 원본서열을 기준으로 부호화할 대상서열을 정렬하는 단계;

상기 원본서열과 상기 대상서열의 차이점을 비교하는 단계;

상기 정렬된 원본서열과 대상서열을 소정의 크기로 분할하는 단계;

상기 원본서열과 대상서열의 차이점을 소정 개수의 문자에 의해 문자열로 변환하는 단계;

상기 문자열을 구성하는 각각의 문자를 소정 개수의 변환코드에 의해 코딩하는 단계;를 포함하는 것을 특징으로 하는 DNA 서열 부호화 방법.

【청구항 8】

제 7항에 있어서,

상기 문자는 DNA를 구성하는 염기를 나타내는 4개의 제1문자, 0 에서 9까지의 9개의 제2문자, 상기 차이점의 시작 및 종료를 나타내는 1개의 제3문자, 및 상기 차이점의 연속여부를 나타내는 1개의 제4문자로 구성되는 것을 특징으로 하는 DNA 서열 부호화 방법.

【청구항 9】

제 8항에 있어서,

상기 변환단계는,

상기 차이점 각각에 대해 상기 차이점의 시작을 나타내는 상기 제3문자를 부여하는 단계;

상기 차이점의 시작위치를 나타내는 상기 제2문자를 부여하는 단계;

상기 차이점의 연속여부를 나타내는 상기 제4문자를 부여하는 단계;

상기 차이점을 구성하는 염기가 연속되는 베이스의 개수를 나타내는 상기 제2문자를 부여하는 단계;

상기 차이점을 구성하는 염기를 나타내는 상기 제1문자를 부여하는 단계;

상기 차이점의 종료를 나타내는 상기 제3문자를 부여하는 단계;

상기 차이점의 시작위치로부터 상기 차이점의 종료위치까지의 거리를 나타내는 상기 제2문자를 부여하는 단계; 및

상기 부여된 문자가 연속적으로 배열된 상기 문자열을 출력하는 단계;를 포함하는 것을 특징으로 하는 DNA 서열 부호화 방법.

【청구항 10】

제 7항에 있어서,

상기 차이점의 형태는 상기 원본서열과 상기 대상서열의 시작영역이 불일치하는 시작영역불일치, 상기 원본서열에는 존재하는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하지 않음을 나타내는 공백, 상기 원본서열과 상기 대상서열의 대응되는 하나

의 베이스 위치에 상이한 염기가 존재하는 단일베이스쌍불일치, 상기 원본서열에는 존재하지 않는 염기가 상기 대상서열의 대응되는 베이스 위치에 존재하는 삽입, 상기 원본서열과 상기 대상서열의 대응되는 복수의 베이스 위치에 상이한 염기가 존재하는 다중베이스쌍불일치, 및 상기 원본서열과 상기 대상서열의 종료영역이 불일치하는 종료영역불일치를 포함하는 것을 특징으로 하는 DNA 서열 부호화 방법.

【청구항 11】

제 7항에 있어서,

상기 변환코드는 상기 소정 개수의 문자 각각에 대응되는 4비트의 코드인 것을 특징으로 하는 DNA 서열 부호화 방법.

【청구항 12】

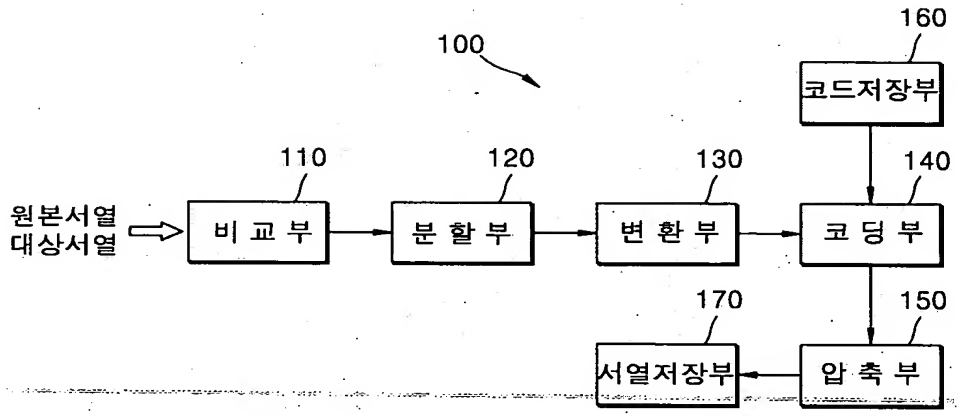
제 7항에 있어서,

상기 변환코드에 의해 코딩된 상기 대상서열을 압축하는 단계; 및

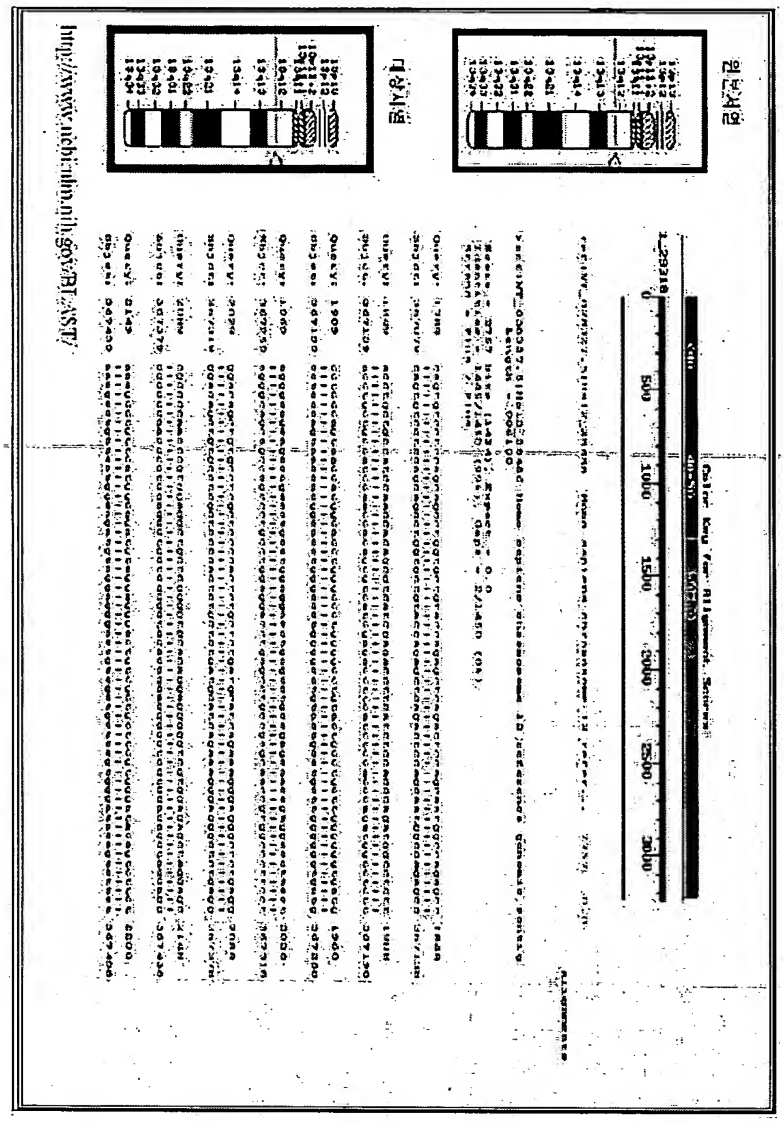
상기 압축된 대상서열을 저장하는 단계;를 더 포함하는 것을 특징으로 하는 DNA 서열 부호화 방법.

【도면】

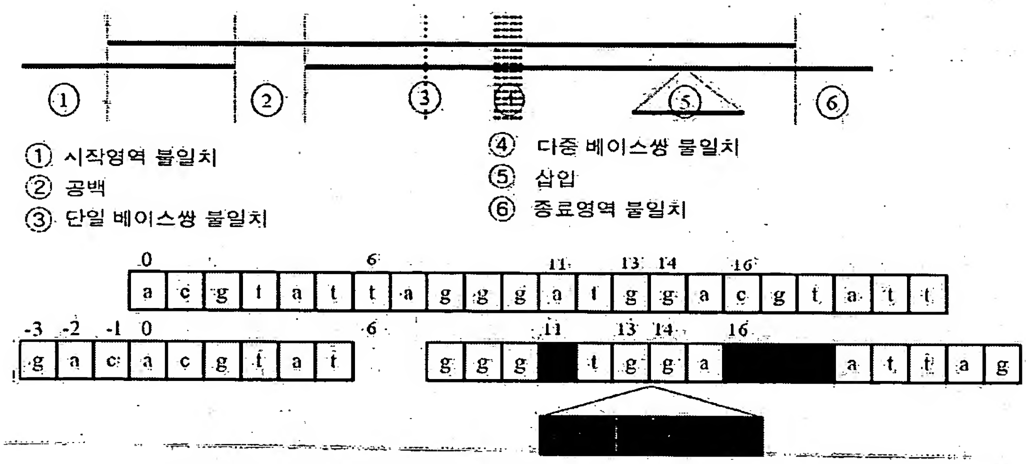
【도 1】


100
160
코드저장부
110
120
130
140
원본서열
대상서열
비 교 부
분 할 부
변 환 부
코 딩 부
170
150
서열저장부
압 축 부

【도 2】



【도 3】



/-3~3gac/3/6/2/3~1c/1/1~6atgcat/1/2~3tcc/3/4~2ag/2

【도 4】

DNA 서열의 차이점을 4bit로 부호화 하기 위한 코드	
0: 0000	8: 1000
1: 0001	9: 1001
2: 0010	A: 1010
3: 0011	T: 1011
4: 0100	G: 1100
5: 0101	C: 1101
6: 0110	/ : 1110
7: 0111	~ : 1111

【도 5】

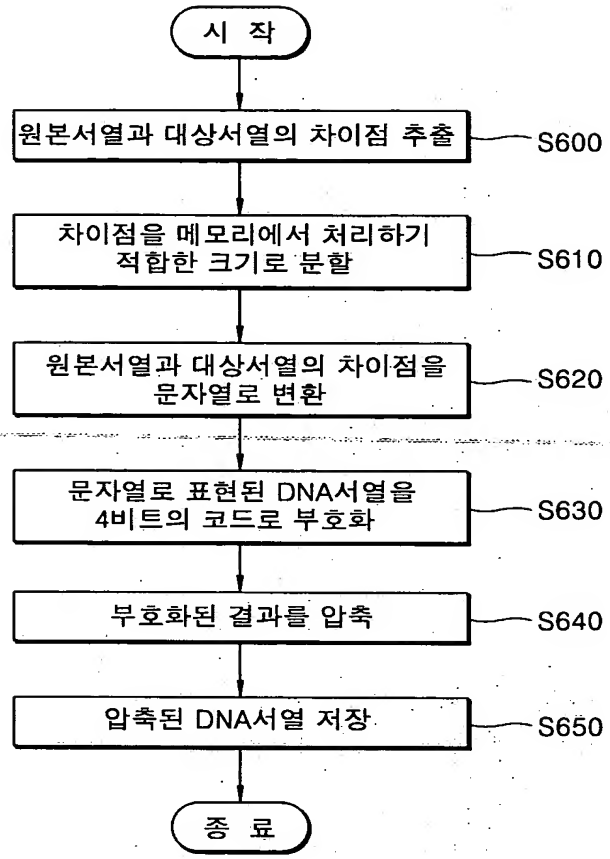
MODY3 exons (5552 characters = 5552 byte)
ref|NT_028327.5|Hs12_28486 Homo sapiens chromosome 12 reference genomic contig
Length = 906100

Exon2 /354759-1a/1
Exon3 /360197-1t/1
Exon456 /362119-1a/1
Exon456 /362178-1a/1
Exon789 /363208-1t/1
Exon789 /363265-1t/1
Exon789 /363261-1t/1
Exon789 /363267-1t/1
Exon789 /363273-1t/1
Exon789 /363278-1t/1
Exon789 /363384-1t/1
Exon789 /363301-1t/1
Exon789 /363394-1t/1
Exon789 /363641-1a/1
Exon789 /363680-1a/1

/354759-1a/1/4438-1a/1/2922-1a/1/59-1a/1/30-1a/53-1t/1/4-1t/1/2-1/t/1/6-1t/1/5-1t/1/6
1t/1/7-1t/1/3-1t/1/247-1a/1/248-1a/1
122byte

1110001101010100011101011001111000011010111000011110010001000011100011100001101011100001111000101001001000010
1110000110101110000111100101100111100001101011100001111000110000111000011010111001010011111000011011111100001
1110010011100001101111100001111000101110000111101011111000011110011011100001101111100001111001011110000110111
1110000111100110111000011011111000011110011111100001101111100001111000111110000110111110000111100010010001111
11100001101011100001111000100100100011100001101011100001
61byte

【도 6】


시 작
원본서열과 대상서열의 차이점 추출
S600
차이점을 메모리에서 처리하기 적합한 크기로 분할
S610
원본서열과 대상서열의 차이점을 문자열로 변환
S620
문자열로 표현된 DNA서열을 4비트의 코드로 부호화
S630
부호화된 결과를 압축
S640
압축된 DNA서열 저장
S650
종 료